Financial Topic Models

Gabriel Doyle
Department of Linguistics
University of California, San Diego
La Jolla, CA 92093-0108
gdoyle@ling.ucsd.edu

Charles Elkan
Department of Computer Science
University of California, San Diego
La Jolla, CA 92093-0404
elkan@cs.ucsd.edu

Abstract

We apply topic models to financial data to obtain a more accurate view of economic networks than that supplied by traditional economic statistics. The learned topic models can serve as a substitute for or a complement to more complicated network analysis. Initial results on S&P500 stock market data show that topic models are able to obtain meaningful stock categories from unsupervised data and show promise in revealing network-like statistics about the stock market. We also discuss the characteristics of an ideal topic model for financial data.

1 Introduction

The financial crisis of the last year has served as a stark reminder of the intricate webs that comprise our modern economy, and the critical need to better understand their structure. Traditional economic analysis has regarded the economy as an overly simplified system where large entities drive the economy, and smaller entities are either ignored or considered only in aggregate. However, the current crisis was caused not merely by the failure of a few large entities, but by a confederacy of problems, including the collapse of the housing bubble, the credit crunch, and the very structure of the market [1]. Network-based approaches [2, 3] have been used to create more accurate analyses of economic systems, but these have their own shortcomings. Notably, they suffer from a lack of sufficient, dependable, and relevant data. We argue that topic models of economic systems can serve as a bridge between traditional economics and network theory, resulting in a more accurate view of the economic landscape than that of traditional methods, while avoiding or mitigating the data problems of network analysis.

In this paper, we look at the specific case of modelling the structure of the stock market, using stock price fluctuation data to infer topics over various publicly-traded companies. These topics are based on what companies exhibit consistent co-movements in stock price, suggesting that they are connected in an economically relevant manner. Each topic can be thought of as representing a sector in the economy, and because of the nature of topics, companies can appear in more than one sector, and companies can be distinguished by their importance within a topic. These topics can be used in various ways to discover the layout of the economy. First, we show that topic models classify stocks into meaningful sectors. Second, we give examples of what topic models can reveal about the nature of these sectors within the economy. Lastly, we discuss how topic models can be used to supply weights for network models.

2 Financial topics

Our goal with the stock market task is to find groups of stocks that tend to move together. An obvious example is that companies within the same industrial sector might be expected to rise or fall together. Collaboration between companies or presence on the same supply chain might also tie companies' stock prices together. However, co-moving stocks need not move in the same direction. Sectors that

Table 1: Four topics learned from the S&P 500 data. These topics had the highest average probability ($\bar{\theta_z}$) across documents of appearing out of the 100 learned topics.

Topic 1 $\bar{\theta_z} = .021$	Topic 2 $\bar{\theta_z} = .019$	Topic 3 $\bar{\theta_z} = .017$	Topic 4 $\bar{\theta_z} = .017$
Southwestern Energy	Penneys	Capital One	Simon Property
Range Resources	Macys	BNY Mellon	Kimco Realty
Cabot Oil & Gas	Kohls	Discover	Equity Residential
EOG Resources	Nordstrom	Northern Trust	AvalonBay Communities
Chesapeake Energy	Target	Janus	Apartment Investment
Pioneer Resources	Limited	JPMorgan Chase	Vornado Realty Trust
Devon Energy	Lowes	State Street	Boston Properties
Peabody Energy	Home Depot	Wells Fargo	Public Storage
Anadarko Petroleum	American Express	PPL	Host Hotels
Massey Energy	Abercrombie	T. Rowe Price	HCP Inc.

are in competition may consistently move together, but in different directions. A financial topic is then a collection of companies whose stock values move together. These topics can correspond to industrial sectors or subsectors, groups of competing firms, or groups of companies with even subtler ties. Unlike network methods, which generally derive connections between companies through loans, trading, or other collaborative ties, topic models should be able to infer connections even when the companies do not have direct ties.

To build the topic model, we begin by viewing the stock market as a collection of price changes divided up into days. On analogy to text processing, each day of stock trading can be viewed as a single document. On each day, the price of a stock may rise or fall by some percent. To get discrete data, we round to the nearest whole percent. "Words" in the documents are stock symbols with a direction of change, either positive or negative. A stock that falls N percent in a single day will have N copies of its negative word in that day's document, and similarly a stock that rises N percent will have N copies of the positive word. For instance, if on Monday Apple's stock (*AAPL*) falls 3% and Google's (*GOOG*) rises 2%, then the Monday document would be *AAPL– AAPL– AAPL– GOOG+ GOOG+*.

We then use standard topic model methods to learn topics over these words. We assume a generative model where each symbol-direction word w is generated from some topic z. Each topic can loosely be thought of as the set of stocks whose prices will change for a certain reason. A stock that rises 3% might rise 2% due to sector-wide improvement, and 1% due to the improvement of its supply-chain. Topics are distributions $\phi_{w,z} = p(w|z)$ over words, where $\phi_{w,z}$ is the probability that a given stock change w would result from a topic z. The probability of a topic z causing a price fluctuation on a given day d is given by $\theta_{z,d} = p(z|d)$. (Note that the ϕ and θ distributions are analogous to the ϕ and θ distributions in text modelling.)

3 Results

To test the viability of a topic model on financial data, we implemented a simple unsupervised LDA model [4], trained on stock price changes from the S&P 500. Our dataset consists of 501 days worth of trading data from January 2007 to September 2008, with 469642 total "words" (symbol-direction pairs). We constructed a model with 100 topics. Samples of topics, with the ten most likely symbols for each topic, are shown in Table 1.

These topics are immediately interpretable. Topic 1 is a set of energy companies, connected by the commercial development of shale-gas in Marcellus Shale Formation. Topic 2 covers mall and big-box retail stores, Topic 3 spans the major financial firms in the S&P500, and Topic 4 covers major residential and commercial real-estate developers. These topics fit with the Global Industry Classification Standard (GICS) system, which classifies the S&P500 stocks into 10 sectors. Each topic's top ten companies share the same GICS sector classification. In fact, the topic model divides companies further than the GICS, as it separates topics 3 and 4, which are lumped into the same "Financials" sector in GICS. Thus we see that topic models can unsupervisedly extract meaningful classifications from stock market data, showing that topic models are appropriate for financial data.

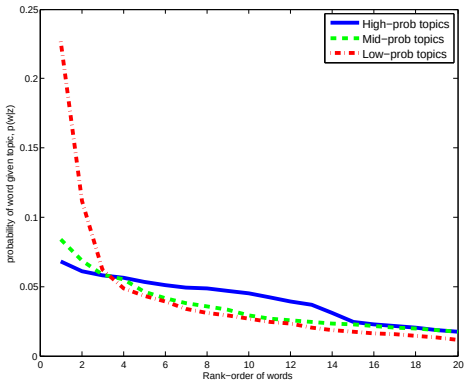

Figure 1: Rank-order plot of $p(w|z)$ values for the most prominent words in three sets of topics: the ten most likely to appear in a document ($\bar{\theta} = .0168$), ten of moderate probability ($\bar{\theta} = .0091$), and the ten least likely to appear ($\bar{\theta} = .0059$). Each line is the average of the ten topics. More peaked lines indicate narrowly-focused topics, less peaked lines indicate broad topics.

In addition to the information gleaned from topic membership, we can investigate the relative importance of each topic to see what drives the market. A topic's importance can be estimated by its average probability, $\bar{\theta_z} = \frac{1}{D}\sum_d \theta_{z,d}$. This importance is the measure used in Table 1 to identify the most important topics in the model. Different topics may be either focused, affecting only a few stocks, or broad, affecting large swaths of the market.

Comparing focused and broad topics in terms of their $\bar{\theta_z}$ can estimate the relative importance of general versus specific events in changing stock prices. Figure 1 shows that the most prominent topics are fairly broad, while less prominent topics are more focused. The x-axis is the rank order of words in a topic, ranked by $p(w|z)$, and the y-axis is the value of $p(w|z)$. The lines are the average values of $p(w|z)$ over likely topics (blue line), unlikely topics (red dot-dash) and middle probability topics (green dashed). This suggests that sector-wide factors have the strongest effects on the market and more specific factors take a back seat.

It is also interesting to note that even the broadest topics are fairly focused; there does not seem to be a general "all stocks go up" topic as we might have expected. This gives us some notion about the structure of the market. It also suggests that, for the S&P 500, there is no wide-ranging interconnectedness between companies; this is an interesting and perhaps soothing result, because high interconnectedness can increase the systemic risk of failure cascades in an economic system [1]. However, this low interconnectedness may be due to the S&P 500's composition, as it contains only a few companies from each of a large set of industrial sectors. As such, a more complete set of stocks should be investigated to properly estimate interconnectedness.

3.1 Building networks from topics

A topic model can also be used to construct a network, using the topic model to estimate the weights of the connections between companies in the market. Analyses of economic networks has risen in prominence because they can explain economic phenomena that traditional economic indicators cannot. Recent work by Reyes et al [3] illustrates this point by comparing the growth of high-performing Asian economies to the stagnation of Latin American economies. Traditional economic indicators, such as per capita GDP, fail to explain this disparity, but calculating the countries' centrality in the global trade network reveals that the Asian economies' growth is correlated with increased network centrality. Network analysis has also yielded insights into the landscape of various economic arenas, such as commercialization strategies amongst biotech firms [5]. Knowledge of the network topology is crucial in determining a network's susceptibility to feedback processes, and the systemic risk of failure cascades.

Despite its strength, network analysis is often hampered by data issues. Determining a network's topology requires high-quality and comprehensive data that is often difficult to obtain, especially in intriguing sectors such as financials. Even when the data is available, it may not be appropriate for identifying patterns or quantifying influence within the network [1]. Part of the problem is that economic network models are intended to be influence networks, but the available data may only

be weakly correlated with influence. Another part of this problem is that the data commonly used to build networks, such as debt-credit relationships, address only individual components of a company's overall state. By building a topic model using stock price, which functions as a noisy holistic indicator of a company's overall status, we learn topics that show which companies behave similarly. Building a network based on similarity of response to events should be effective at tracking the possibility of cascading failures and other systemic risks.

The strength of a connection between two companies can be easily estimated once a topic model has been trained. Each company w has a vector $\vec{\Phi_w} = (\phi_{w,1}, \cdots, \phi_{w,K})$, where $\phi_{w,i}$ is the probability that a change in the price of company w will be generated by topic i. These vectors define unnormalized probability vectors that can be compared across companies. Two companies with similar normalized $\vec{\Phi_w}$ vectors appear in the same topics with similar probabilities, and thus will be expected to have some sort of close ties, whether collaborative or competitive. Because this measure is based on the holistic metric of market prices, we expect it to be at least as good, if not better, at capturing network topology than connections based on a single non-holistic measure.

4 Expanding financial topic models

In the above experiments, we used a basic LDA implementation. To fully realize the potential of a financial topic model, we want a model with three traits: the ability to learn how many topics to use, the ability to modify topics over time, and the ability to handle bursty data.

A model that learns the appropriate number of topics is useful, as we are hoping to find some unexpected topics and we have no reasonable way of knowing how many such topics to expect. To satisfy this characteristic, a topic model like hierarchical LDA [6] can be used. Furthermore, we expect that topics will change both in composition and prominence over time. Once-vibrant sectors will shrink, and nascent sectors will grow. Companies will enter and leave markets, creating new competitions, and enter and leave supply chains. Thus it is essential that the topics be dynamic, and able to adjust to the passage of time. A dynamic topic model [7] will satisfy this criterion. Lastly, financial data is inherently bursty. Burstiness refers to a situation where a word appearing once makes it more likely to appear again, yielding a leptokurtic distribution; that financial data follows such a distribution has been known for more than fifty years [8]. DCMLDA [9] has been shown to account for this burstiness in learning a topic model.

Unfortunately, no topic model has yet been proposed that simultaneously satisfies all three of these criteria. Financial topic models can get started by considering a model that uses just one of these criteria, but for the best results, a hybrid model is needed.

References

[1] F. Schweitzer, G. Fagiolo, D. Sornette, F. Vega-Redondo, A. Vespignani, and D. R. White. Economic networks: The new challenges. *Science*, 325:422–425, 2009.

[2] M. Serrano and M. Boguñá. Topology of the world trade web. *Physical Review E*, 68, 2003.

[3] J. Reyes, S. Schiavo, and G. Fagiolo. Assessing the evolution of international economic integration using random walk betweenness centrality: The cases of East Asia and Latin America. *Advances in Complex Systems*, 11(5):685–702, 2008.

[4] D. Blei, A. Ng, and M. Jordan. Latent Dirichlet allocation. *J. of Machine Learning Research*, 3:993–1022, 2003.

[5] W. Powell, D. White, K. Koput, and J. Owen-Smith. Network dynamics and field evolution: The growth of interorganizational collaboration in the life sciences. *American J. of Sociology*, 110:1132–1205, 2005.

[6] D. Blei, T. Griffiths, M. Jordan, and J. Tenenbaum. Hierarchical topic models and the nested Chinese restaurant process. In *Advances in Neural Information Processing Systems 16*, 2004.

[7] D. Blei and J. Lafferty. Dynamic topic models. In *Proceedings of 23th International Conference on Machine Learning*, 2006.

[8] M. G. Kendall. The analysis of economic time-series. *J. of the Royal Statistical Society. Series A (General)*, 116:11–34, 1953.

[9] G. Doyle and C. Elkan. Accounting for burstiness in topic models. In *Proceedings of 26th International Conference on Machine Learning*, 2009.